DOCUMENT RESUME

ED 362 040 FL 021 504

AUTHOR Shohamy, Elana
TITLE The Power of Tests: The Impact of Language Tests on Teaching and Learning. NFLC Occasional Papers.
INSTITUTION Johns Hopkins Univ., Washington, DC. National Foreign Language Center.
PUB DATE Jun 93
NOTE 23p.
AVAILABLE FROM National Foreign Language Center, 1619 Massachusetts Ave., Washington, DC 20036 ($5 prepaid).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Arabic; *Construct Validity; Elementary Secondary Education; English (Second Language); *Language Tests; National Competency Tests; Oral Language; Reading Comprehension; Second Language Learning; *Testing; *Testing Problems; Uncommonly Taught Languages

ABSTRACT

This paper is rooted in an expanded view of construct validity, whereby the role of testers does not end in the development phase of the language tests they employ. Rather, testers need to follow the uses of these tests and examine issues of utility, relevance, ethics, and interpretation. The studies reported here focused on three national language tests, and examined their impact on teaching and learning in the school context. The three tests were: a test of Arabic as a second language for seventh, eighth, and ninth grades; an English-as-a-Second-Language oral test; and first-language reading comprehension test for fourth- and fifth-grade students. Data were collected through class observation, questionnaires, interviews, and analyses of documents. The impact of all three tests was complex, occurring in a number of directions, and dependent on the nature and purpose of the test. All tests diverted attention to areas that had not been explicitly taught previously. In terms of the test effect, in all three cases, instruction became testlike. Other findings involved tests as de facto curriculum, conflict between teachers and bureaucrats with regard to the use of test results, and the use of tests for purposes different from those that were initially intended. (Contains 14 references.) (JP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 362 040

The Power of Tests: The Impact of Language Tests on Teaching and Learning

ELANA SHOHAMY
Tel Aviv University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard Lambert
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

The National Foreign Language Center
Washington DC

BEST COPY AVAILABLE

FL021504

About the Occasional Papers

This is one in a series of Occasional Papers published by the National Foreign Language Center. The NFLC prints and distributes articles on a wide variety of topics related to foreign language research, education, and policy. The Occasional Papers are intended to serve as a vehicle of communication and a stimulant for discussion among groups concerned with these topic areas. The views expressed in these papers are the responsibility of the authors and do not necessarily reflect the views of the NFLC or of the Johns Hopkins University.

About the National Foreign Language Center

The National Foreign Language Center, a nonprofit organization established within the Johns Hopkins University in 1987 with support from major private foundations, is dedicated to improving the foreign language competency of Americans. The NFLC emphasizes the formulation of public policy to make our language teaching systems responsive to national needs. Its primary tools in carrying out this objective are:

- *Surveys.* The NFLC conducts surveys to collect previously unavailable information on issues concerning national strength and productivity in foreign language instruction, and our foreign language needs in the service of the economic, diplomatic, and security interests of the nation.
- *National policy planning groups.* In order to address major foreign language policy issues, the NFLC convenes national planning groups that bring together users of foreign language services and representatives of the language instructional delivery systems in formal education, the government, and the for-profit sector.
- *Research.* The NFLC conducts research on innovative, primarily individual-oriented strategies of language learning to meet the nation's foreign language needs of the future.

In addition, the NFLC maintains an Institute of Advanced Studies where individual scholars work on projects of their own choosing.

The results of these surveys, discussions, and research are made available through the NFLC's publications, such as these Occasional Papers, and they form the basis of fresh policy recommendations addressed to national leaders and decision makers.

Copyright © 1993, The National Foreign Language Center at the Johns Hopkins University

Price per copy: $5.00, prepaid

BEST COPY AVAILABLE

The Power of Tests: The Impact of Language Tests on Teaching and Learning

Elana Shohamy
Tel Aviv University

Few devices are as powerful, or are capable of dictating as many decisions, as tests. As Madaus (1990) puts it, "A single standardized test score independently triggers an automatic admission, promotion, placement or graduation decision" (p. 5). This paper reports the results of three studies that examined the uses and impact of language tests within the educational context in which they operate. Thus, this paper is rooted in a broader view of construct validity, one that claims that construct validity encompasses aspects of test use, the impact of tests on test takers and teachers, the interpretation of scores by decision makers, and the misuses, abuses, and unintended uses of tests (Messick 1981, 1989).

AN EXPANDED VIEW OF CONSTRUCT VALIDATION

According to this broader view of construct validity, the role of testers does not end in the development phase, when they achieve high reliability and validity; rather, they need to follow the uses of these tests and examine issues of utility, relevance, ethics, and interpretation. Testers can no longer be viewed as technicians whose work is done when they reach satisfactory reliability coefficients; rather, they must consider the social, psychological, ethical, curricular, and educational consequences of the tests they produce. As Messick (1981) points out,

> The test user cannot be the sole arbiter of the ethics of assessment, because the value of measurement is as much a scientific and professional issue as the

Elana Shohamy is a professor of language education at the School of Education, Tel Aviv University. Her research and publications focus on the area of language testing and address such issues as the connection between language testing and learning, diagnostic testing, oral testing, the effects of contextual variables on test scores, and social and ethical aspects of language tests. She is the author, with Herbert W. Seliger, of *Second Language Research Methods* (Oxford, 1989) and the editor, with A. Ronald Walton, of *Language Assessment for Feedback: Testing and Other Strategies* (Kendall/Hunt and the National Foreign Language Center, 1992). During 1989–90 and the summers of 1991 and 1992 she was a Mellon fellow at the National Foreign Language Center. This paper was presented at an NFLC colloquium on September 14, 1992.

The author acknowledges the contribution of the following graduate students in the collection and analysis of data: Zahava Ganor, Amalia Haskal, Dalia Peltz, and Ola Peri (the Arabic test); Perle Arie, Malka Ashkenazi, and Ronit Wiser (the EFL oral test); and Sharon Edit, Bracha Lev, Dalia Segal, Amira Rom, and Hani Shilton (the reading comprehension test).

> meaning of measurement. One implication of this stance is that the published research and interpretive test literature should not merely provide the interpreter with some facts and concepts but with an accounting of the critical value contexts in which those facts are embedded and with a provisional reckoning of the potential social consequences of alternative uses. (p. 19)

The need to include aspects of test use in construct validation originates in the fact that testing is not an isolated event; rather, it is connected to a whole set of variables that interact in the educational process. Results obtained from tests have serious consequences for individuals as well as for programs, since many crucial decisions are made on the basis of test results. Among these are the placement of students in class levels, the granting of certificates or diplomas, determinations as to whether students are capable of continuing in future studies, the selection of students most suitable for higher-education institutions, and the acceptance of job applicants and program candidates.

Foucault (1979) describes tests as the most powerful and efficient tool through which society imposes discipline. In *Discipline and Punish* he points to the specific features that enable tests to become so powerful—the ability to observe, perform surveillance, quantify, classify, normalize, judge, and punish:

> The examination combines the techniques of an observing hierarchy and those of a normalizing judgement. It is a normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish. It establishes over individuals a visibility through which one differentiates them and judges them. That is why, in all the mechanisms of discipline, the examination is highly ritualized. In it are combined the ceremony of power and the form of the experiment, the deployment of force and the establishment of truth. At the heart of the procedures of discipline, it manifests the subjection of those who are perceived as objects and the objectification of those who are subjected. (pp. 184–85)

The power and authority of tests enable policymakers to use them as effective tools for controlling educational systems and prescribing the behavior of those who are affected by their results—administrators, teachers, and students. This phenomenon can be observed in a variety of settings and contexts. Policymakers in central agencies, aware of the authoritative power of tests, use them to manipulate educational systems, to control curricula, and to impose new textbooks and new teaching methods. At the school level, principals use tests to drive teachers to teach, and teachers use tests to force students to study. Schoolwide exams are used by principals and administrators to enforce learning, while in classrooms tests and quizzes are used by teachers to impose discipline and to motivate learning (Stiggins and Faires-Conklin 1992). Thus, tests initially designed to provide information on achievement or to select suitable candidates for jobs have become devices for controlling and manipulating educational systems.

The use of tests for power and control is an especially common practice in countries that have centralized educational systems, where the curriculum is controlled by central agencies (Shohamy 1991). In such countries tests are viewed as the primary tools through which changes in the educational system can be

introduced without having to change other educational components such as teachers training or curricula. In such systems it is understood that the introduction of national tests will trigger additional factors affecting the educational process. Tests such as matriculation examinations or national achievement surveys are used as an attempt to cure complex educational malaises.

Consider the example of a national supervisor of language who had been criticized because students graduating from high school were found to be lacking in reading comprehension and writing skills. To "cure" this complex problem, he decided to change the end-of-high-school national matriculation examination by adding reading comprehension and writing components. All students graduating from high school in that country would have to demonstrate their competence in those two areas. The supervisor therefore had to find ways for them to practice reading comprehension and writing. This led to the production of new books "guaranteeing" success on the new test, and students made an intensive effort to study an area they had never before been explicitly taught, in preparation for a test that would have a significant effect on their lives (e.g., gain them entrance to a university). Interestingly, reading comprehension and writing had always been included on the national curriculum, but since these areas had not been tested on the national exam, teachers did not take them seriously. The introduction of the new test implied that these areas were indeed important: only if something is tested is it important.

Consider yet another example—a new EFL (English as a foreign language) oral test battery introduced nationally as part of the end-of-high-school exam. The rationale given by the national supervisor was that only the introduction of the new test could ensure that oral language would be taught in EFL classrooms; because of the test, teachers would increase the amount of time they devoted to teaching oral language.

In these two examples tests were not only used to manipulate and control education; they also became the devices through which educational priorities were communicated to principals, teachers, and students.

Although this phenomenon is more typical of centralized than of decentralized educational systems, it is interesting to observe that the latter, though they do not have national curricula, also rely on tests for controlling and driving education. In the United States, for example, there is currently a trend toward introducing tests in order to drive instruction, and a new national test is being proposed to upgrade the level of learning (Madaus 1990). Typical of the drive to introduce national end-of-high-school exams is the belief that they will "cure" the low levels of achievement among American students that were reported in international surveys.

Language tests in the United States have also played a major role in the effort to drive and improve foreign language learning. This effort, led by the American Council on the Teaching of Foreign Languages, was implemented through the ACTFL Proficiency Guidelines and the Oral Proficiency Interview (OPI); the guidelines were aimed at providing a description of criteria and standards of

proficiency, while the OPI was offered as the testing procedure through which proficiency would be assessed. One reason for introducing these instruments was to put pressure on teachers and students to upgrade the level of foreign language learning.

The connection between testing and learning is demonstrated by a number of newly introduced terms: *washback effect* refers to the impact tests have on teaching and learning; *measurement-driven instruction* refers to the notion that tests should drive learning; *curriculum alignment* focuses on the connection between testing and the teaching syllabus; and *systemic validity* implies the integration of tests into the educational system and the need to demonstrate that the introduction of a new test can improve learning. Still, while the connection between testing and learning is commonly made, it is not known whether it really exists and, if it does, what the nature of its effect is. Can the introduction of tests per se cause real improvement in learning and teaching? And how are test results used by teachers, students, and administrators?

Such questions have been addressed in a number of studies recently reported (Stake 1989; Wall and Alderson 1992; Xiaoju 1990; Smith 1991; Shepard 1991). Stake (1989) studied the effect of introducing new statewide tests as devices for upgrading achievement and found that although the tests caused teachers to become more focused in their teaching, they had the additional effect of narrowing the scope of the subject matter taught. Smith (1991), examining the effect of external tests on teachers, found that teachers have negative feelings regarding the publication of test scores and will do whatever is necessary to avoid low scores. She showed that in the classroom, the need to prepare for an external test substantially reduces the amount of time available for instruction; preparation for the test is not always compatible with teaching. Xiaoju (1990) reported on a study that examined the impact of a new language test in China, the MET (matriculation English test), and found that the test encouraged the use of new textbooks and innovative teaching materials.

The studies reported here focused on three newly introduced national language tests, and examined their impact on teaching and learning in the school context—specifically, what happened to teaching once a new test was introduced. The three tests were a test of Arabic as a second language for seventh, eighth, and ninth grades; an EFL oral test, part of the national tests administered to twelfth-grade students at the end of secondary school; and an L1 reading comprehension test for fourth- and fifth-grade students. The data was collected through observations of classes, questionnaires, interviews, and analyses of documents. The design, data-collection procedures, and results for each of the tests are described below.

1. THE ARABIC TEST

The Arabic test was introduced by the Ministry of Education for seventh-, eighth-, and ninth-grade students learning Arabic as a second language. The test was first

administered in 1988 for seventh-grade students only; it is currently administered in the middle of each school year to all three grade levels. The Arabic curriculum for the seventh grade includes the teaching of the alphabet, grammatical structures, and vocabulary of about three hundred words; the eighth- and ninth-grade curricula both add about three hundred new words a year.

The main reason for introducing the test, according to Arabic inspectors at the Ministry of Education, was to raise the prestige of the Arabic language, to compare the level of teaching of Arabic in schools throughout the country, to motivate teachers to speed up the teaching of Arabic, and to increase the motivation of both teachers and students. Clearly, the Ministry of Education felt that the study of Arabic needed a "push" and that the test could be instrumental in providing it.

One goal for introducing the seventh-grade test was to reduce the amount of time teachers spent on teaching the alphabet. Previously it had taken about two years, but the national inspector of Arabic believed that this ought to be reduced to six months. The administration of the test was therefore scheduled for six months after the beginning of the school year. The seventh-grade test included items that required letter recognition, vocabulary, and grammar. In the first few years of the test, the inspector distributed a vocabulary list from which the test vocabulary was selected.

Three research questions were posed regarding the impact of the Arabic test:

- How did the introduction of the test affect teaching practices?
- How did the introduction of the test affect students' behavior?
- What was the long-range impact of the test?

How Did the Introduction of the Test Affect Teaching Practices?

To answer the first question, data was collected through lesson observations, a review of teaching materials, interviews with teachers, and student questionnaires. Questionnaires were collected on two separate occasions: one to four weeks before the test, and one to two weeks following it. Observations were conducted in nine seventh-, eighth-, and ninth-grade Arabic classes taught by different teachers. Each class was observed three times: twice before the test, and once following it. Each observation was followed by an interview with the teacher and an analysis of the teaching material. The number of students in each class ranged from thirty-nine to forty-two. The focus of the observation was teaching activities associated with the test.

Teaching activities before the administration of the test. Observations prior to the administration of the test revealed that class time and activities were dominated by it; there were constant references to "the test," and it had become the central focus and goal of all classroom activities. Typical of these lessons were review and preparation sessions with clear goals in terms of outcome. Specifically, only material known to be on the test was discussed; any deviation from this was

considered a diversion and a waste of time. The teachers were fully aware of this and explained that if they continued with "regular" teaching, they would not have enough time to prepare their students for the test, and the students would fail. The review period lasted between four and six weeks.

The following activities were typical of the preparation and review period:

- Teachers stopped teaching new material and turned to reviewing material.
- Teachers stopped using the class textbooks, in the belief that the textbooks did not provide a good basis for the review because they did not cover the same material that would be included on the test.
- Worksheets that were essentially "clones" of the previous year's test were produced by teachers for use during the review period. These included long word lists, up to three hundred words long, with translations (for all classes) and the alphabet (for the seventh grade); some actually included previous versions of the test.
- There was extensive use of Hebrew in the classroom; teachers claimed that using Arabic was a waste of time.
- The activities were all "testlike"—that is, the material was identical in format to what was expected to be on the test.
- The main class activities included grammar reviews, mechanical drills, translation of words, practice of the alphabet, and memorization of new words, based on word lists prepared and distributed by the Ministry of Education.
- Vocabulary was taught in a decontextualized manner, using word lists accompanied by Hebrew translations.
- There was ample use of repetition, rote practice, and mechanical grammar exercises.
- The main homework activities were "more of the same"—activities mostly designed to allow students to complete at home what could not be covered during class time. Most of these activities were clones of the test, assigned after the teacher had explained the material briefly in class.
- Review sessions of up to two hours a week were added to regular class hours.
- Learning of new material took place only when it was needed for the test, and then it was done at a rushed pace.
- Tests were used extensively as devices for teaching material that had not been taught before (see example below).
- There was no error correction and very little explanation of material. There were no checks of whether the students had internalized or mastered the new material, or of whether they knew it in different contexts; it was mostly mechanical rote learning. There were quick shifts from one topic to another; the impression was that quantity must come at the expense of quality, since there was so much to cover in so little time.
- The atmosphere in the class was tense, on the part of both students and

teacher (see example below). Yet students were highly motivated to master the material, and no discipline problems were recorded. Students felt that there were clear goals for learning.

Teaching activities after the administration of the test. Observations conducted after the test showed no references at all to "the test"; it was back to regular teaching. The following were typical of the activities that took place after the test had been administered:

- New material was taught.
- Teaching was done through textbooks.
- Vocabulary was taught in a contextualized way, via stories and conversations.
- Language was taught communicatively, through teachers' personal stories, discussions of pictures, and so on.
- There was little homework, and it was varied and contextualized.
- There was use of L2 in the classroom.
- The class atmosphere was relaxed; the pace was slower and seemed less efficient. There were some discipline problems.
- No tests were used throughout the whole observation period in any of the classes.

Table 1 summarizes the differences between teaching activities before and after the test.

The following example illustrates some of the activities and the atmosphere in one class before and after the test. The description is based on three observations.

Observation 1 (before the test): In the lesson that preceded this one, the teacher had distributed a vocabulary list of three hundred words with their Hebrew translations. These words had been selected from the list of words for the test,

Table 1: Comparison of Activities before and after the Arabic Test

Before the Test	**After the Test**
review of material	teaching of new material
use of worksheets	use of textbooks
isolated vocabulary lists	contextualized vocabulary (stories, conversations)
decontextualized teaching	communicative/contextualized language
rote learning, memorization, drills	contextual, meaningful activities
ample homework	little homework
use of L1	use of L2
frequent use of tests for teaching	no use of tests for teaching
tense atmosphere	relaxed atmosphere
rushed pace	slow pace
no discipline problems	some discipline problems

and the students had been asked to memorize them at home. However, the students came to class unprepared and had not mastered those words. The teacher was angry, the students complained that it was too much to learn at one time, and the teacher complained that the students were lazy. The teacher started to test individual students orally, to prove that they did not know the words. As expected, the students did not know the words, and they were upset and tense. The teacher gave them an assignment to prepare for a test on the same vocabulary for the next lesson; the test would include one hundred of the words. The students had to memorize the words on their own; no teaching of vocabulary took place.

Observation 2 (five days later, before the test): The teacher returned the tests that had been administered in the previous lesson (not observed). The teacher stated that 90 percent of the students had failed the test, and she accused the students of not studying hard enough. She tested them again, orally, by inviting individual students to the board to correct words that were wrong on the test. The teacher dictated the words and asked students to translate them as part of class work while she went around checking each student's work.

Observation 3 (one week after the test): The atmosphere was relaxed, the teacher was calm. Students were asked to open their books to a new story. The teacher read new vocabulary lists, then she read a text that included the words and held a discussion on the content of the story utilizing the new words. This activity lasted until the end of the session. The teacher did not translate any of the new words (of which there were twenty); there was frequent use of L2. However, not all the students participated, some were bored, there were some discipline problems, and the teacher admitted later (in the interview) that there had been a lack of attention and a decrease in motivation since the administration of the test.

In summary, the examination of the effect of the test on teaching practices showed that there was a sharp distinction between teaching and testing: teaching stopped, and test review began. There was constant reference to the test, and all activities were geared to it. No new material was taught during the review period, except for topics needed for the test. There was ample use of worksheets, homework, and tests as preparation devices. The main goal was to succeed on the test. After the test was administered, learning of new material resumed. Teaching consisted of contextualized and communicative activities and the use of L2. Compared with the period before the test, the lessons were less focused and less efficient.

How Did the Introduction of the Test Affect Students' Behavior?

Data on the effect of the test on student behavior was collected through questionnaires administered to forty-five students and from the above-mentioned classroom observations.

While 45 percent of the students claimed that the test had not affected them, 55 percent claimed that it had. Those who claimed to have been affected said that because of the test, they

— listened more carefully during lessons;
— paid more attention to the subject;
— took Arabic studies more seriously;
— were more highly motivated to learn; and
— obtained private tutoring to make up for missed material.

In terms of the type of impact, 62 percent claimed that the test had affected them positively, while 38 percent claimed that it had affected them negatively. Those who claimed to have been affected positively said that the test

— forced them to learn more Arabic;
— enriched their knowledge of the subject matter and improved their grades;
— helped them master new vocabulary;
— enriched their language; and
— motivated learning.

Those who claimed to have been affected negatively said that the test

— induced fear, pressure, and anxiety;
— frustrated them, since they felt that the test did not reflect real learning;
— gave them a feeling of wasted time; and
— did not improve their proficiency in Arabic.

Thus, different students were affected by the test differently. For some it was beneficial, for others it was not. All students claimed that they learned more because the material was more focused. They felt that they had a better idea of what was expected from them in learning Arabic.

What Was the Long-Range Impact of the Test?

The long-range impact of the test was examined through data collected after the test had been given for four years in its present format. The data collection for this question focused on the seventh grade only, since for that grade there was an explicit goal of shortening the amount of time required to teach the Arabic alphabet. Thus, one type of impact examined was the extent to which that goal was achieved. Other questions that were examined related to whether the teachers used the test content as the teaching content; whether the test changed teaching methods; whether it changed teachers' perceptions of Arabic as a school subject; and whether it changed the status of Arabic as a school subject.

The data for examining the long-range impact of the test was obtained from twelve teachers of Arabic from various parts of the country whose seventh-grade classes were tested. They filled in questionnaires, personal interviews were held, and their classes were observed. The questionnaire consisted of sixteen questions that addressed aspects of the test's impact on teaching methods and the perception of Arabic as a school subject. Four of the questions focused on the teachers'

backgrounds, six on teaching and testing methods, and six on perceptions of Arabic as a school subject. In addition, interviews were conducted with each of the teachers, to validate the responses obtained from the questionnaires. Classroom observations were aimed at finding out whether there were overt or covert references to the test and to examine the specific activities teachers were engaged in before the administration of the test. In addition, students' notebooks were reviewed.

The results of the questionnaire showed that half of the teachers said that they had been affected by the test and that it had influenced their teaching, while the other half said that they had not been influenced; however, as it turned out, all those who claimed not to have been influenced were new teachers who had recently graduated from teacher-training institutions where they had been trained in teaching the alphabet in a shorter amount of time. All the teachers who claimed to have been influenced by the test had been teaching for more than five years.

Those who claimed to have been influenced by the test said that it

- gave them direction as to the setting of new teaching priorities;
- affected their allocation of time, causing them to reduce the amount of time devoted to teaching the alphabet;
- gave them direction as to what aspects of Arabic needed to be taught, and how;
- upgraded the status of the Arabic subject (only if something is tested is it considered to be important); and
- created pressure and tension, and interrupted regular teaching.

The results of classroom observations performed before and after the test showed that there was very little special preparation for the test, since the new textbooks being used were a direct reflection of the test. The most apparent impact of the test was the development of new textbooks based on new approaches to teaching the alphabet and including the same activities that were on the test. Thus, while in the early years of the test there was a distinct difference between teaching and testing, that difference subsided in the latter period as new textbooks that reflected the test appeared on the market. Today the alphabet is taught in a substantially shorter amount of time, and teaching includes the vocabulary included on the test. Thus, over the years teachers have begun to teach Arabic in the "new way," geared to the test. The gap between testing and teaching has therefore diminished as the two have become integrated. The new textbooks have become, de facto, the new curriculum.

In conclusion, when the new Arabic test was first introduced, there was a strong and immediate test effect as teaching stopped and test preparation began. The material needed for the test became the body of knowledge to be mastered. The preparation for that body of knowledge was instrumental, and both students and teachers knew what they wanted to achieve. Activities became very focused and efficient, covering only material known to be on the test. Thus, learning

became narrow, mechanical, and superficial; it was expressed through "testlike" activities like worksheets that replicated the test in terms of format and content, review materials, special lessons preparing students for the test, and a large dose of testing and quizzes. The strong impact of the test lasted for about four to six weeks; afterwards, teaching returned to the nontest format. However, as time went by, teaching and testing become synonymous. This was mainly a result of the introduction of new textbooks designed to match the new body of knowledge, the test material. These textbooks were clones of the test, especially in terms of their activities and tasks; they became the new teaching material. Testing and teaching became quite similar, very few review sessions were observed, and teachers admitted to having been influenced by the test in terms of direction and guidance. The Ministry of Education was therefore successful in introducing new material through the device of a test.

2. THE EFL ORAL TEST

The EFL oral test is part of the national matriculation examination administered at the end of twelfth grade to all students graduating from high school. A new oral test was introduced in 1986 after a series of experiments (Shohamy, Reves, and Bejarano 1986). It consists of a number of tasks representing the following speech interactions: oral interviews, role-plays, reporting or picture descriptions, and literature tests. The test is administered to individual students in two separate stations by different testers. The first station consists of an oral interview and a role-play; the second consists of a picture description and a literature test. Each station lasts about seven minutes. The oral test that preceded the new one consisted of an unstructured oral interview that was similar to a conversation. The main rationale for introducing the new oral test was to increase the emphasis on oral language in the EFL classroom, and thus to upgrade the speaking proficiency of students.

This study examined the impact of the new oral test on the teaching of spoken English in the classroom. The data was collected from observations and interviews with fifteen teachers—ten teachers with up to five years' experience, and five new teachers. The main findings are described below.

Time spent on oral language. All the experienced teachers claimed that the test affected their behavior so that they now spent substantially more time on oral language in the classroom than they did before the new test was introduced. Five of the teachers claimed to have spent no time at all on oral language before the test was introduced. The novice teachers—those who had taught for up to three years—claimed that the test did not affect their behavior, since they had been trained in the teaching of oral language at their teacher-training institutions.

Classroom activities. As to the specific class activities used to teach oral language, it was found that they were identical to the activities included on the test; the same tasks used for testing oral language were used as teaching activities—class interviews, role-plays, and picture descriptions and reports. No other

activities were observed or reported in the interviews. This tendency may have been reinforced by a booklet published by the Ministry of Education in preparation for the test, which included lists of topics and role-plays. This booklet was designed to ease the transition from the old to the new oral testing system. It provided a source from which teachers could draw specific activities for practicing for the test.

Perceptions of oral language. With regard to how teachers perceived oral language, the findings showed that they perceived it exclusively in terms of testlike activities. Thus, when asked to define "oral language," teachers often gave answers such as "It is a role-play" or "It is an interview."

Experienced vs. novice teachers. There were major differences between experienced and novice teachers with regard to the influence of the test. The new teachers tended to try out more untestlike activities in the teaching of oral language. For example, they undertook a variety of communicative activities in their classes such as debates, simulations, lectures, plays, discussions, and group activities. They claimed that having this type of test actually opened new teaching avenues to them, and they therefore experimented with innovative types of interaction. The difference between novice and experienced teachers is probably due to the fact that the novice teachers had been trained in oral language teaching in their teacher-training programs and were therefore familiar with a variety of methods for teaching oral language. The experienced teachers, on the other hand, who had not obtained such training, turned to the test as their main source of guidance for teaching oral language; they viewed the oral test as an additional burden they had to deal with in order to prepare their students for the test, and they took the shortest possible route to that goal.

3. THE READING COMPREHENSION TEST

The L1 reading comprehension test was administered nationally to all students in the fourth and fifth grades by the Ministry of Education in the spring of 1991. The test consisted of sixty questions: fifteen of the questions were defined as the minimal level required in reading comprehension proficiency (answering two of the fifteen questions incorrectly resulted in a fail), and the rest constituted a norm-referenced test in which the score was reported as the percentage of correct answers. The format of the test was short passages followed by questions pertaining to the text and the vocabulary included in it.

The administration of the test was accompanied by a great deal of media attention, and strong resentment on the part of the teachers. The results were disseminated in a special news conference, in which it was reported that 33 percent of the students had failed the test. A map of the country was shown on national television, highlighting cities and areas where large numbers of students had failed. It was a major news event, with about five hundred newspaper articles being written on the subject. The test had political, social, economic, and educational implications.

The data on the impact of the test was collected from interviews with teachers and from an examination of materials produced after the test was administered. The results reported below describe the impact on specific areas.

Teaching materials. After the release of the results, ample new teaching material was produced by teachers and regional supervisors; about thirty books and workbooks were published. Careful examination of the materials showed that they were mostly clones of the test in terms of format—texts plus questions (usually multiple-choice ones). Supervisors developed practice pages and worksheets identical to those used on the test.

Allocation of time. Many teaching sessions were diverted to the teaching of reading comprehension, so that the number of hours allocated to the subject increased. This is important, because previously no special hours had been allocated to reading comprehension; rather, it had been integrated into other subjects. Now content areas such as geography, history, and the like were turned into reading comprehension sessions.

Teachers. On the basis of interviews with ten teachers, five being teachers whose classes had failed, the most obvious impact was found to concern emotional involvement and stress. Teachers whose classes had failed did not want to be identified by name. They spoke endlessly about the test—indeed, they were happy to have the opportunity to do so; they expressed anger and frustration, and were very critical of the test. They claimed that they had been wrongly blamed for their classes' failure; the fault lay with the student population. The teachers complained that the test did not reflect the material they taught; that the format did not reflect their views of reading comprehension; that they had not been told in advance what would be tested; and that the unfamiliar format must have increased students' anxiety and affected their performance. The teachers were humiliated by the fact that they had not been specifically consulted about the test so that it would reflect the teaching that was taking place. They felt that principals were blaming them for the results; that an external body was interfering and intruding into their privacy; and that without being consulted, and without understanding why, they were suddenly being told to go in different directions, as if they had been wrong all along. They also felt that they should have had a choice as to *who* got tested.

Test results. Because no guidance was provided regarding the use of the results, there was extensive misuse of them. Teachers were blamed for the students' failure by both principals and parents. Principals used the results to justify their own opinions of teachers. This was the case with one teacher who, because she felt it suited her students, was using a teaching method different from the one her principal had recommended. She claimed that her class failed not for any reason having to do with the teaching method, but because there were five new immigrants and five disabled children in her class. But the principal used the results to convict her of having used the wrong method. The teacher was not allowed to teach that class the following year. In the same way, teachers whose classes did well on the test were rewarded.

Teachers and administrators. By and large, principals and other administrators thought that it was a good idea to give the test. Teachers, on the other hand, felt that it was unnecessary. All teachers thought that the test was humiliating and unfair, and that the results simply confirmed what they already knew about their students.

The media and the public. The test had a strong impact in the media; about five hundred articles about the results appeared in different newspapers. Worried about the attacks they were receiving from the public, teachers and teachers' organizations claimed that the test was not valid and did not reflect actual teaching and learning. One editorial in a major newspaper attacked the teachers, contending that they should admit they had failed and not look for excuses. In many schools, posters calling on teachers to ignore the test results were distributed by the national teachers' union. Thus, another effect of the test was to create a negative image of teachers and to engender tension between teachers, the Ministry of Education, and the public.

Public awareness. The test resulted in heightened public awareness of educational issues. In a country where political and military news get most of the attention, an educational issue was for once on the agenda. Reading comprehension became a most talked-about topic, a subject of jokes, major discussions, and debates. The test results were invoked by a number of political parties claiming that "we need peace so we can spend money on reading comprehension, not war." The term *reading comprehension* became familiar to laypeople and was used in everyday conversations.

These additional effects were also observed:

- There was massive production of new materials, worksheets, and textbooks for teaching reading comprehension. Most of these were clones of the test—that is, they offered reading comprehension texts followed by questions, usually multiple-choice ones—and were different from the official curriculum.
- Workshops for teaching reading comprehension were offered around the country, and teachers were asked to attend them.
- Reading was taught entirely in terms of "test activities." Teachers who had a broader view of literacy—those who included writing—turned exclusively to the teaching of reading comprehension, à la test.
- Teachers felt that since the results were not diagnostic, they were not meaningful in terms of pedagogy, and therefore many were at loss regarding what teaching strategies they should use in view of the results.
- Results were used to classify and categorize students in terms of "success" or "failure."

CONCLUSIONS AND DISCUSSION

The following conclusions can be drawn from the findings reported above.

1. All three tests had some type of impact. The impact is complex, occurring in a number of directions, and is strongly dependent on the nature and purpose of the test; it also changes over time.

2. All three tests were instrumental in diverting attention to areas that had not been explicitly taught previously. In the case of the Arabic test, the focus on specific vocabulary and the shortening of the time devoted to the alphabet meant that teaching emphases changed as a result of the test. Similarly, in the case of the EFL oral test, teachers began to devote more time and to give greater emphasis to oral language in the classroom. In the case of the L1 reading comprehension test, teaching began to focus on an area that had not received explicit attention before, although reading comprehension had previously been taught implicitly in every class and in every subject.

   It is interesting to note that in each of these fields, the test content had been included in the curriculum prior to the test, though not in the specific manner the Ministry of Education wanted it to be taught. The Arabic alphabet and vocabulary had been taught, but at a slower pace; English oral language had been taught, but with less emphasis; and L1 reading comprehension had been taught, but through other subjects or through general literacy. Through the introduction of tests, the ministry successfully imposed the teaching of specific topics that it considered important. Students found the test to be helpful in clarifying goals and for learning.

3. In terms of the nature of the test effect, in all three cases the results showed that instruction became testlike. This was found both in teaching methods and in teaching materials. Prior to the introduction of the Arabic test, the focus, method, and pacing of the teaching had been different. After the test, testing materials and methods became an integral part of "normal" teaching as many teaching activities became testlike, mostly as a result of the new textbooks, which were strongly influenced by the test. In the case of the EFL oral test, the teaching of oral language in the classroom became testlike in that it came to involve mainly specific activities and tasks that were included on the test and had not been practiced previously. In the case of the L1 reading comprehension test, all activities and formats became identical to those of the test. The textbooks that emerged in all three areas consist mainly of testlike activities. In all three cases, novice teachers tended to be influenced by the test less than experienced ones in terms of activities and materials, because the test had apparently already had an effect on teacher training as well.

4. In all three cases the use of testlike activities is most likely a result of teachers not having been trained to teach the new areas being tested. Without appropriate pedagogical knowledge, teachers turn to the most immediate and readily available source. In the case of the Arabic test, the new textbooks have become the pedagogical knowledge source, and they are promoted as

such by publishers. Their marketing appeal is that they provide a good source for preparing students for the high-stakes tests, and consequently they consist of test-preparation material. A similar phenomenon has occurred with the other two tests. Thus, in situations where the educational leadership does not provide experienced teachers with on-the-job training in new areas, teachers will turn to the test as their single source of knowledge regarding instruction. One wonders what the expectation of the educational authorities is in terms of teaching, and how they expect teachers to behave if no training in the new area is provided.

5. When teachers receive the message that educational authorities can legitimately impose tests, they tend to do the same themselves, demanding that their students prepare for the test through activities such as memorization, rote practice, and . . . more tests. It becomes a common belief that the act of imposing a test will in itself somehow invoke "learning," and that teachers and students will find ways to cope with the imposition. The instrument used to measure becomes the method used for "therapy."

6. When teaching and testing become synonymous, the tests become the new, de facto curriculum, overshadowing the existing curriculum. This is particularly so in situations where the curriculum is not very explicit and where no instructional guidance is provided by the educational authorities.

7. When the stakes are high, tests are used for purposes different from those that were initially intended. This was observed in the case of the L1 reading comprehension test, where principals used the test scores to judge the teachers, punishing them for students' failure on the test or rewarding them for success. Results are used to frighten, deter, and blame, to justify previous decisions, to impose sanctions, to standardize according to test content, to classify and categorize.

8. Since none of the tests provided detailed diagnostic information, they were not useful for identifying specific weak areas or for offering strategies for repair. Thus, when a class failed, there was no information as to why it had happened or what could be done to improve the situation.

9. In no case were the teachers involved in any way in the preparation of the test. It was imposed on them without their having had any input. The teachers, who should have been viewed as experts in these subjects, became servants of the system, and their authority was challenged.

10. The act of testing does help upgrade the status of the subject being tested, especially when there is high rate of failure, because it invokes strong media exposure and increased public awareness of educational issues.

11. There seems to be a conflict between teachers and bureaucrats with regard to the use of test results. Teachers use results mostly as a source of pedagog-

ical information. Bureaucrats, on the other hand, use tests for different purposes—to serve as tools through which policy can be implemented, to justify previous decisions, to apportion blame, and to prove to the public that action is being taken.

12. The strength of the impact varies, depending on the type of test and on such variables as whether the stakes are high or low, the relevance of the subject to decision makers and to the public at large (certain subjects are considered more important than others), and the rate of failure (a high rate of failure receives more attention from the public and the media).

The following process seems to emerge from the findings on the three test situations: There is an untaught topic or area that the educational leadership decides needs to be taught and mastered. This decision is often a reaction to public or media demands for action. To ensure that the new topic is taught, a (national) test is introduced, since this is the easiest and quickest way for policymakers to demonstrate action and authority. Because of its power and the high stakes, the test serves as an efficient tool for changing the behaviors of teachers and students. Since teachers have not been explicitly trained to teach the new topic, they experience fear and anxiety as students, principals, and parents all demand preparation for this high-stakes test. To overcome the gap in their knowledge , teachers turn to the most immediate pedagogical source—the test itself, or rumors about the material expected to be on it. Over the years new books are written and workshops designed to prepare teachers for the test are given. Thus, if no meaningful professional teacher training takes place, the test becomes the de facto curriculum. Even when a curriculum does exist, it becomes subordinate to the test. In centralized, "authoritative" educational systems, tests become the major device through which the leadership communicates educational priorities to teachers. The teachers, on the other hand, are reduced to simply "following orders"; they are often frustrated by this role, because their responsibility increases while their authority is taken away.

While the introduction of a test can be influential in terms of changing focus, it is not known what the impact could have been if the Ministry of Education had decided to change teaching practices through other means, such as workshops, in-service courses, or new textbooks. The educational effectiveness of tests introduced in such a way cannot be very high, because the approach narrows the process of education, making it merely instrumental and unmeaningful. The test is no more than a quick fix that overlooks the need to attain meaningful comprehension of a subject; no test—especially an external one—can represent more than a limited body of knowledge on any subject. In all three cases discussed here, the tests were used instrumentally, as a fast fix. *Instrumental impact* is characterized as short-range and goal-oriented, while *conceptual impact* is characterized as long-range and meaningful, and is followed by discussions of the nature of the trait, methods of teaching, and so on. In none of the tests was there any serious discussion of what the measured topic meant; rather, quick therapies were

offered as solutions to the problems. The case of the reading comprehension test, in which many schools added reading comprehension hours at the expense of subject areas such as geography and history, rather than integrating it into those areas, is one instance of such simplistic, instrumental solutions. Similarly, the introduction of the Arabic test, in which the complex problem of teaching Arabic was reduced to specific and defined components—the alphabet, vocabulary, and syntax—and *not* communication, provides evidence that bureaucrats are interested in simplistic, instrumental solutions where gains can be seen immediately.

Furthermore, imposing tests without involving teachers—those who are responsible for delivering the instructional information—is a humiliating act based on the view that the role of teachers is to carry out orders, not to initiate, create, or contribute. Teachers view such tests as an intrusion from above that leaves them with responsibility but no authority. It is no surprise that test results tend to be used and misused against teachers.

There are ways of using tests that can also benefit learning. Tests can be used along with many other procedures when new topics are introduced; teachers can be involved in planning and writing such tests; the tests can be geared toward providing diagnostic information and can be used for improvement and repair; teachers can be trained in the new areas that need to be taught; an interactive approach to pedagogy can result in a sharing of both authority and responsibility.

Tests are powerful devices and should be treated as such, but they should not be used as they were by the bureaucrats in the cases described here. Tests are powerful in that they can provide decision makers—students, teachers, and administrators—with valuable information and insights on teaching and learning. It is the information that tests are capable of providing that makes them valuable. For example, information obtained from tests can provide evidence of student ability over a whole range of skills, subskills, and dimensions, and over a whole range of achievement and proficiency, on a continuing basis and in a detailed and diagnostic manner. The information can be used to judge students' language in relation to expectations as outlined in the curriculum; to determine whether the school as a whole is performing well in relation to other schools sharing the same curriculum; to determine whether the teaching methods and textbooks used are effective tools for achieving those goals; and to determine whether the goals are realistic and appropriate. Decision makers need to realize the potential of tests to lead to improvement. Once conclusions are reached on the basis of such information, it is possible to align the curriculum accordingly by implementing changes in teaching methods, textbooks, or expectations. These changes can then be monitored through repeated administration of tests on an ongoing basis.

It is possible to utilize tests by incorporating other factors that are part of the educational process, and to avoid relying on the power of tests per se to create change. Fredericksen and Collins (1989) introduced the notion of *systemic validity* to refer to the introduction of tests along with a whole set of additional variables that are part of the learning and instructional system. Such tests become part of

a dynamic process in which change in the educational system takes place according to feedback obtained from the tests. In systemic models, a valid test is one that brings about, or induces, an improvement in the tested skills after the test has been in the educational system for a period of time. Fredericksen and Collins claim that high systemic validity can be achieved only when a whole set of assessment activities foster it, and they identify a number of such activities.

Tests used for the purpose of improving learning can be effective only if they are connected to the educational system; they are not effective when used in isolation. But using tests to solve educational problems is a simplistic approach to a complex problem. It works on people's fear of authority. It can even be said that the testers themselves are abused by the educational leadership. Testers need to examine the uses that are made of the instruments they so innocently construct.

BIBLIOGRAPHY

American Council on the Teaching of Foreign Languages. 1986. Proficiency Guidelines. Hastings-on-Hudson, N.Y.: ACTFL.

Fredericksen, J., and A. Collins. 1989. "A System Approach to Educational Testing." *Educational Researcher* 18:27–32.

Foucault, M. 1979. *Discipline and Punish.* New York: Vintage Books.

Madaus, G. 1990. "Testing as a Social Technology." Paper presented at the Inaugural Annual Boisi Lecture in Education and Public Policy, Boston College, December 6.

Messick, S. 1981. "Evidence and Ethics in the Evaluation of Tests." *Educational Researcher* 10:9–20.

———. 1989. "Validity." In *Educational Measurement*, ed. R. Linn, pp. 447–74. New York: ACE/Macmillan.

Shepard, L. 1991. "Psychometricians' Beliefs about Learning." *Educational Researcher* 7:2–16.

Smith, M. L. 1991. "Put to Test: The Effects of External Testing on Teachers." *Educational Researcher* 20:8–11.

Shohamy, E. 1991. "International Perspectives of Foreign Language Testing Systems and Policy." In *International Perspectives on Foreign Language Teaching*, ed. G. Ervin, pp. 91–107. Lincolnwood, Ill.: National Textbook Company.

Shohamy, E.; T. Reves; and Y. Bejerano. 1986. "Introducing a New Comprehensive Test of Oral Proficiency." *English Language Teaching Journal* 40:212–20.

Stake, R. 1989. *Effects of Changes in Assessment Policy.* Greenwich, Conn.: JAI Press.

Stiggins, R., and N. Faires-Conklin. 1992. *In Teachers' Hands.* Albany: State University of New York Press.

Xiaoju, L. 1990. "How Powerful Can a Language Test Be? The MET in China." *Journal of Multilingual and Multicultural Development* 11/5:393–404.

Wall, D., and C. Alderson. 1992. "Examining Washback: The Sri Lankan Impact Study." Paper presented at the Fourteenth Language Testing Research Colloquium, Vancouver, February 2.

The National Foreign Language Center

at the Johns Hopkins University
1619 Massachusetts Avenue NW
Washington DC 20036

BEST COPY AVAILABLE